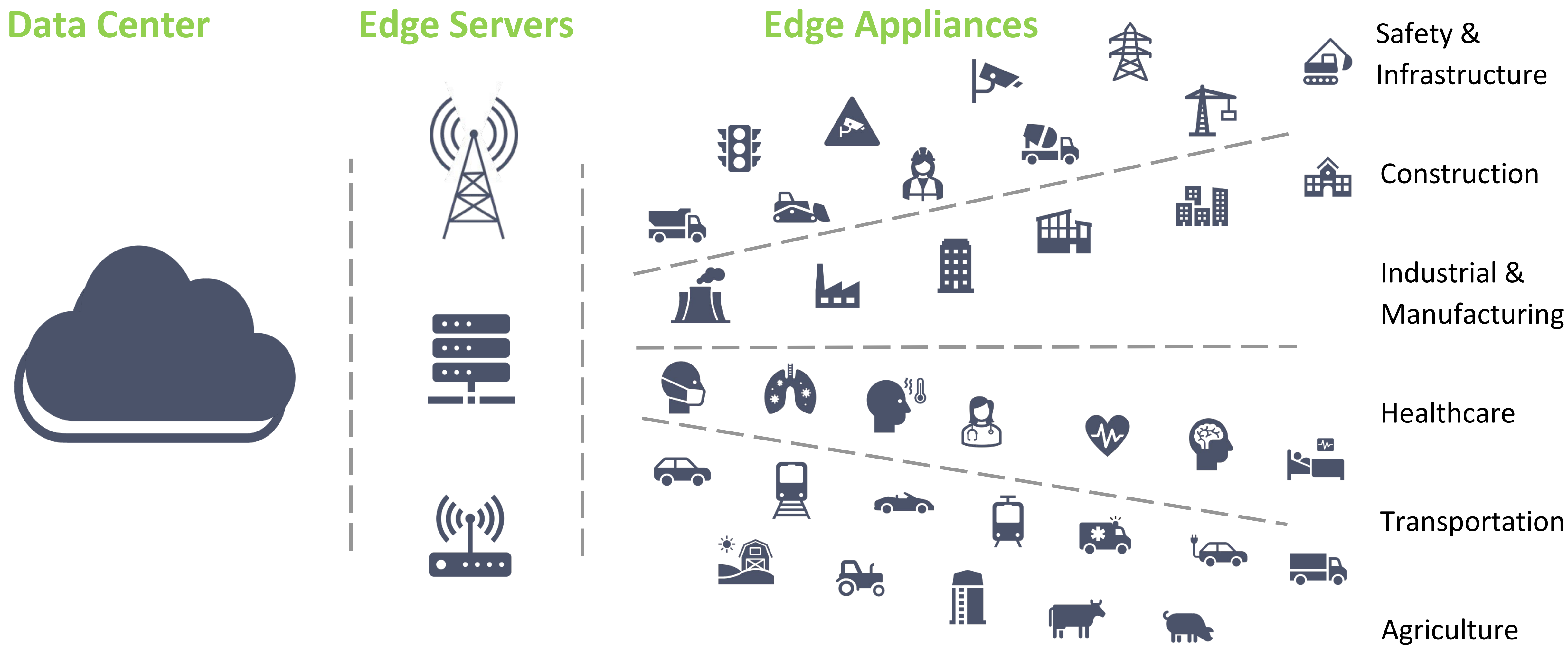
Challenges when Architecting Vision Inference Systems for Transformer Models

AI Hardware Summit

September 2023 | Jeremy Roberson, Director of Inference SW |
jeremy.roberon@flex-logix.com

AI Landscape at the Edge is Dominated by Vision



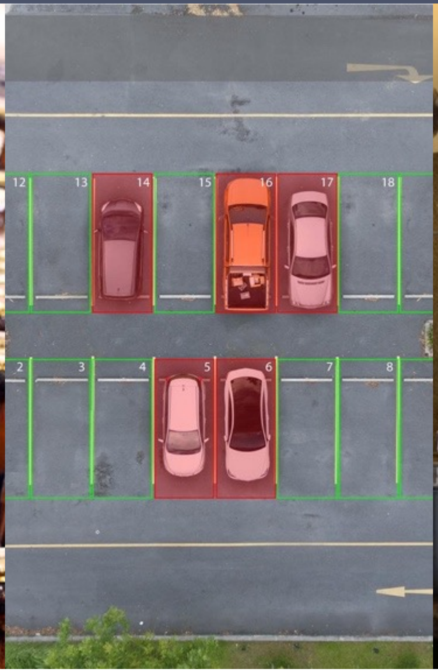
Different requirements in **power** / **cost** / **performance**

Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

Markets and application

Safety & security

Manufacturing & industrial optical inspection



Traffic & parking management

Retail

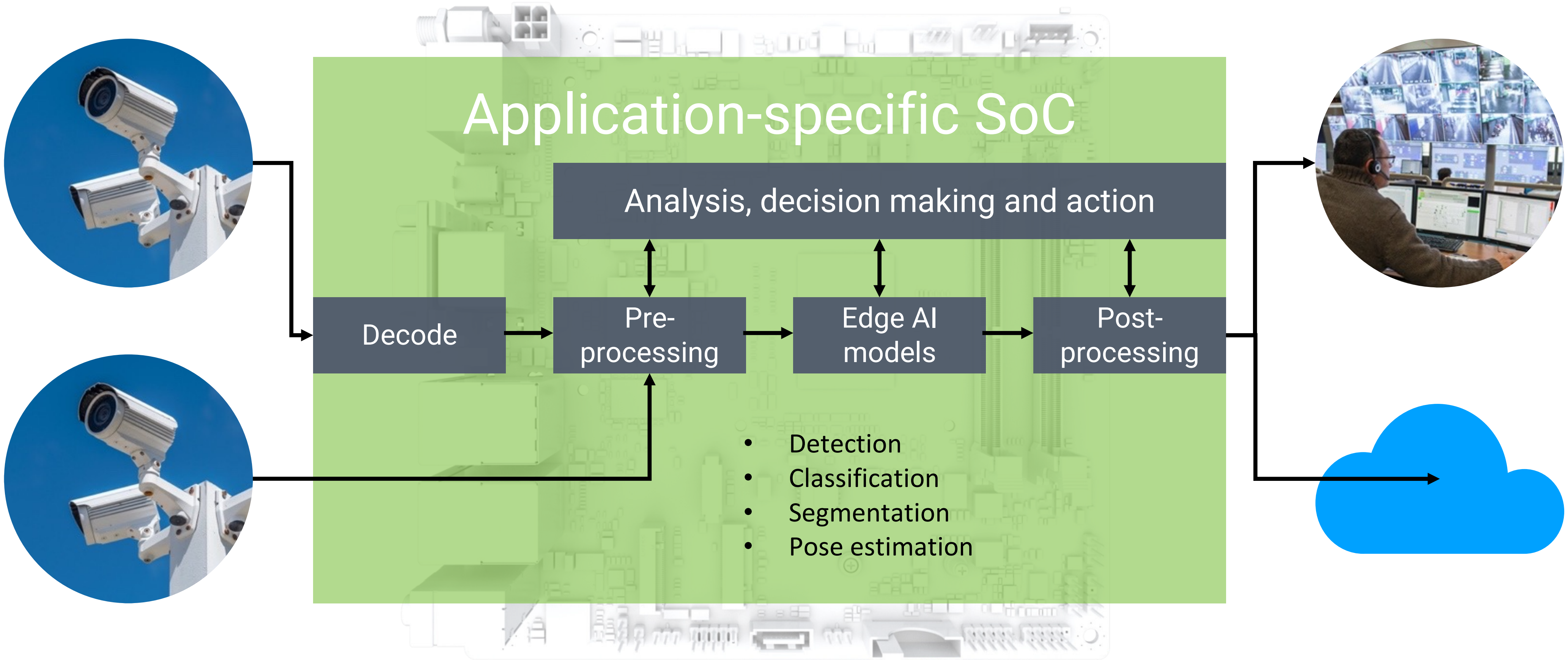
Robotics

Agriculture

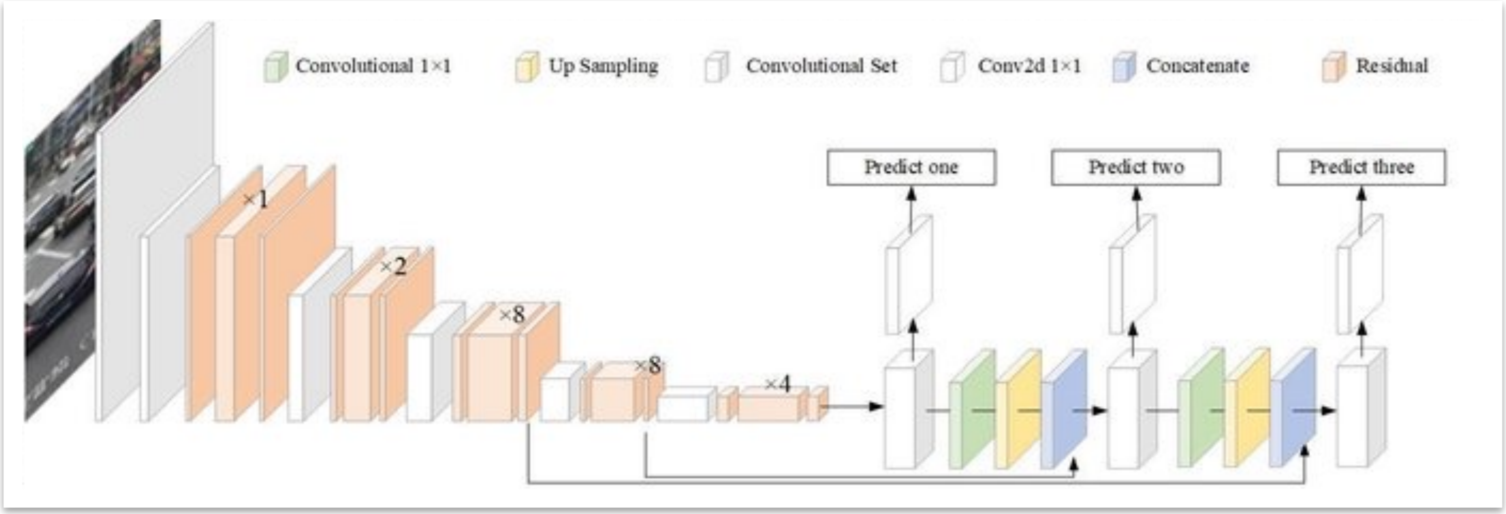
Embedded & Edge SoCs can benefit from AI Computer Vision

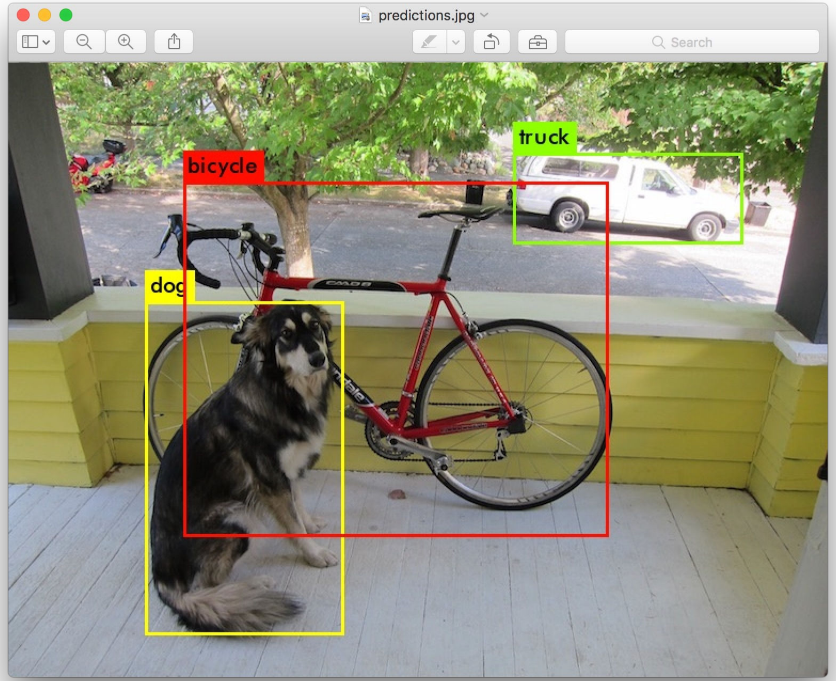
Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

Typical Video and Vision Use cases



Convolutional Networks have fundamentally changed computer vision, but still have limitations



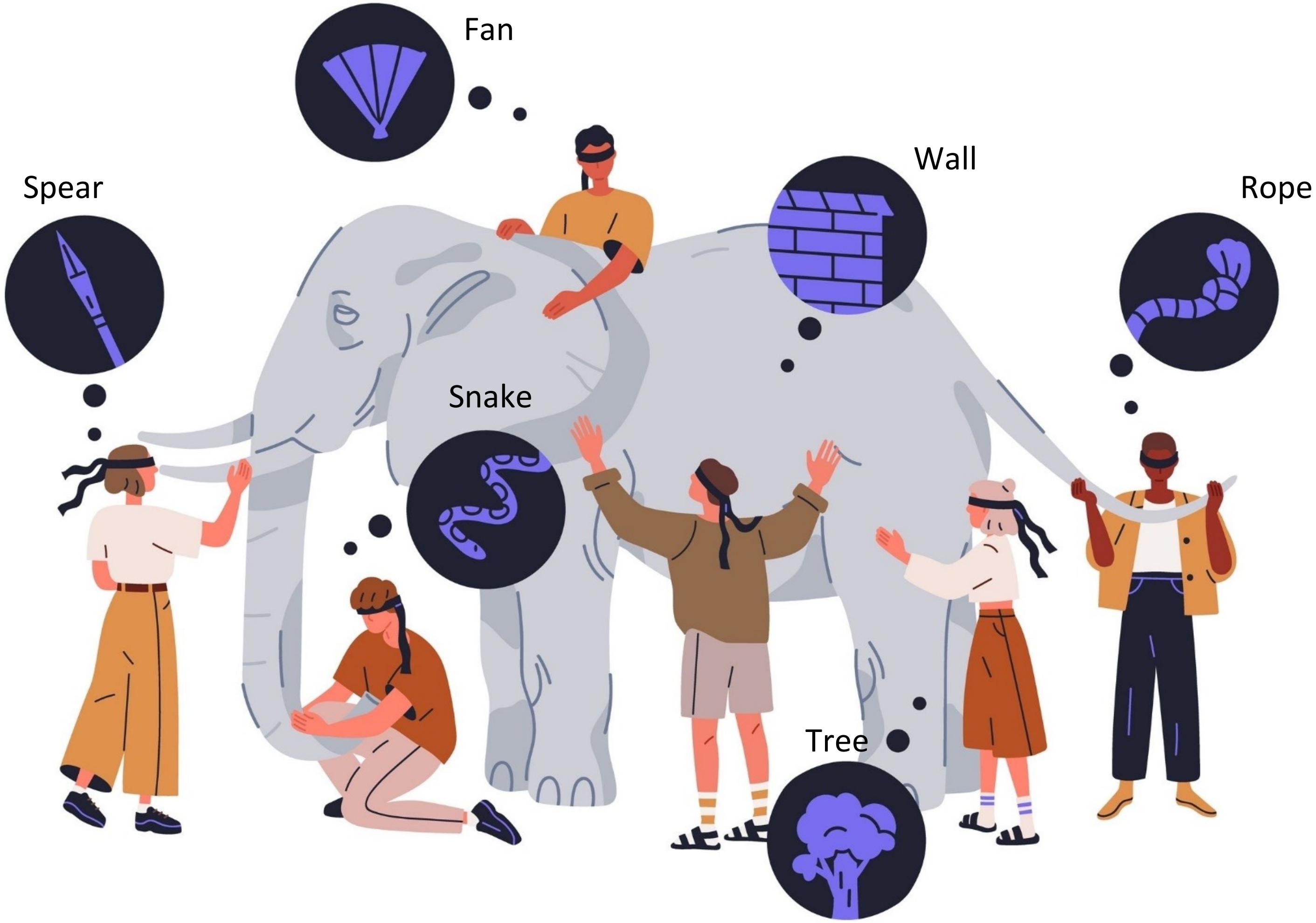


- YOLOv3 example:
- Operator Types: 3x3, 1x1, FC
- Total Layers: 75
- Output is object detection and location for trained categories
- Operations per Inference: 178B at 608x608

flexlogix

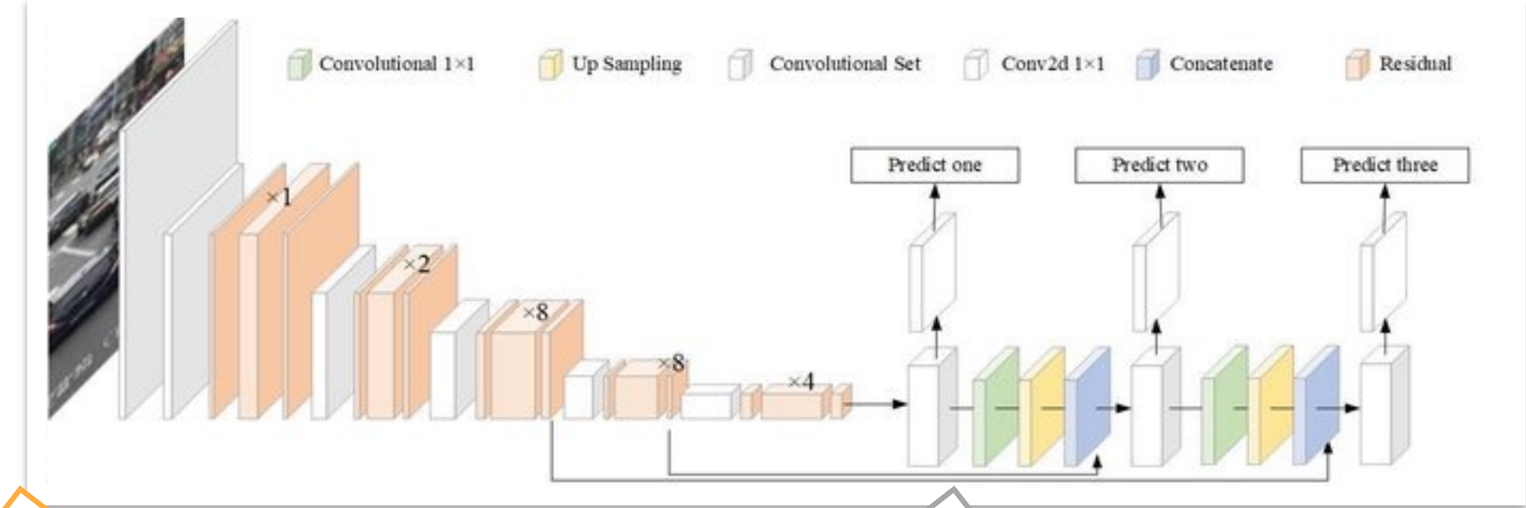
Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

Blind person and the elephant – Receptive Field is a big limitation

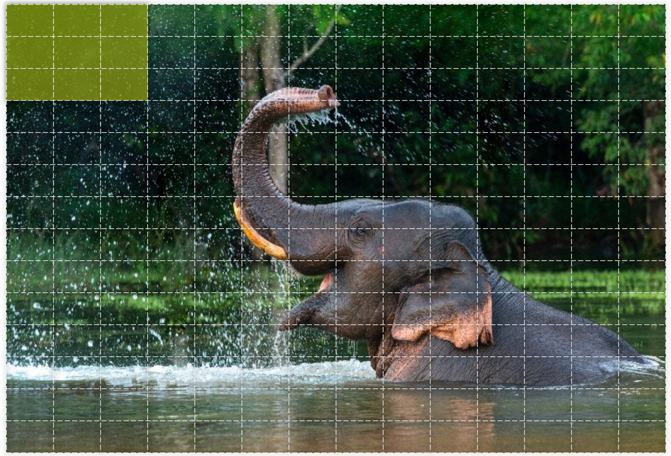


https://en.wikipedia.org/wiki/Blind_men_and_an_elephant#:~:text=The%20parable%20of%20the%20blind,the%20side%20or%20the%20tusk

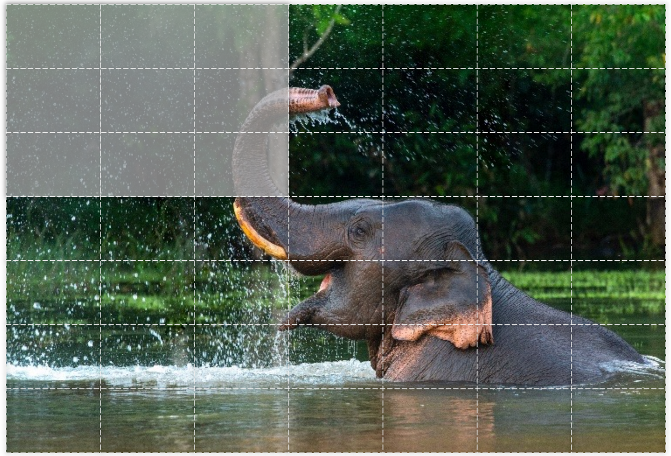
CNN uses larger filter size and down-sampling to increase receptive field, but insufficient



Smaller receptive field
Smaller features & objects



Larger receptive field
Larger features & objects



flexlogix

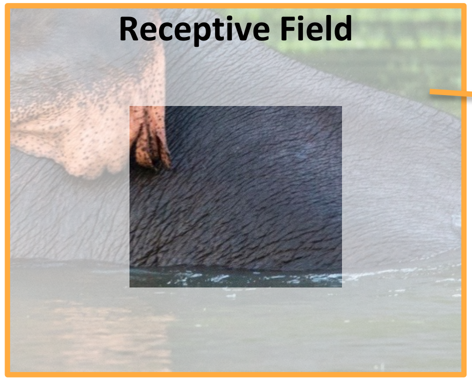
Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

The difference between CNN and transformer models & their feature extraction

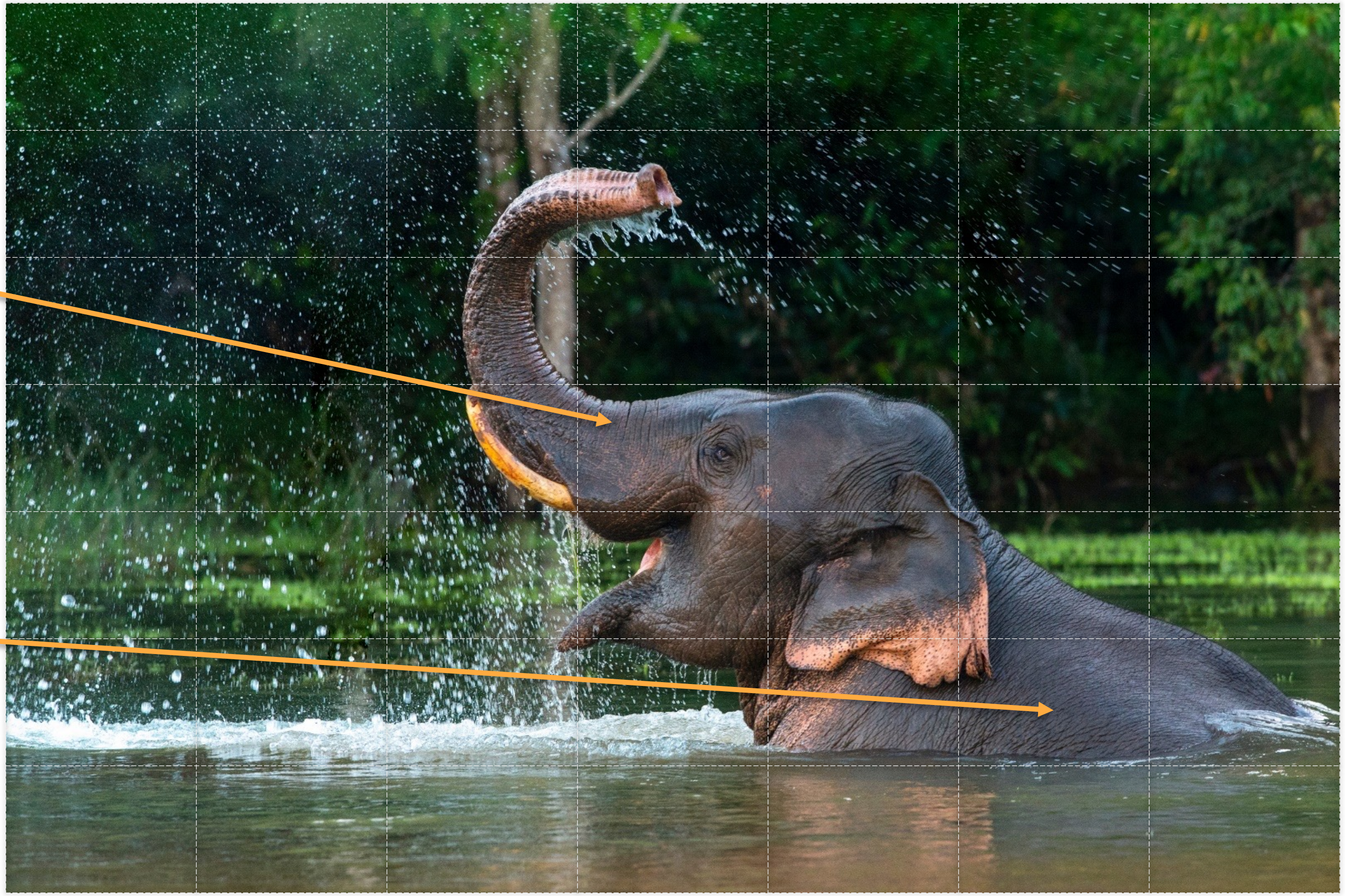
CNNs provide results with limited context

Transformer provide results in context

EASIER

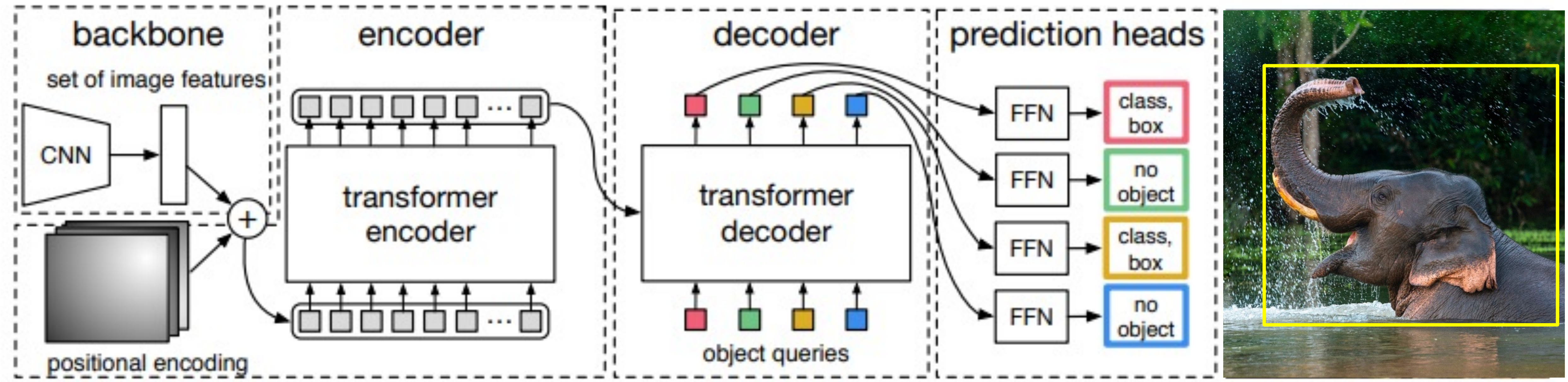


VERY
HARD



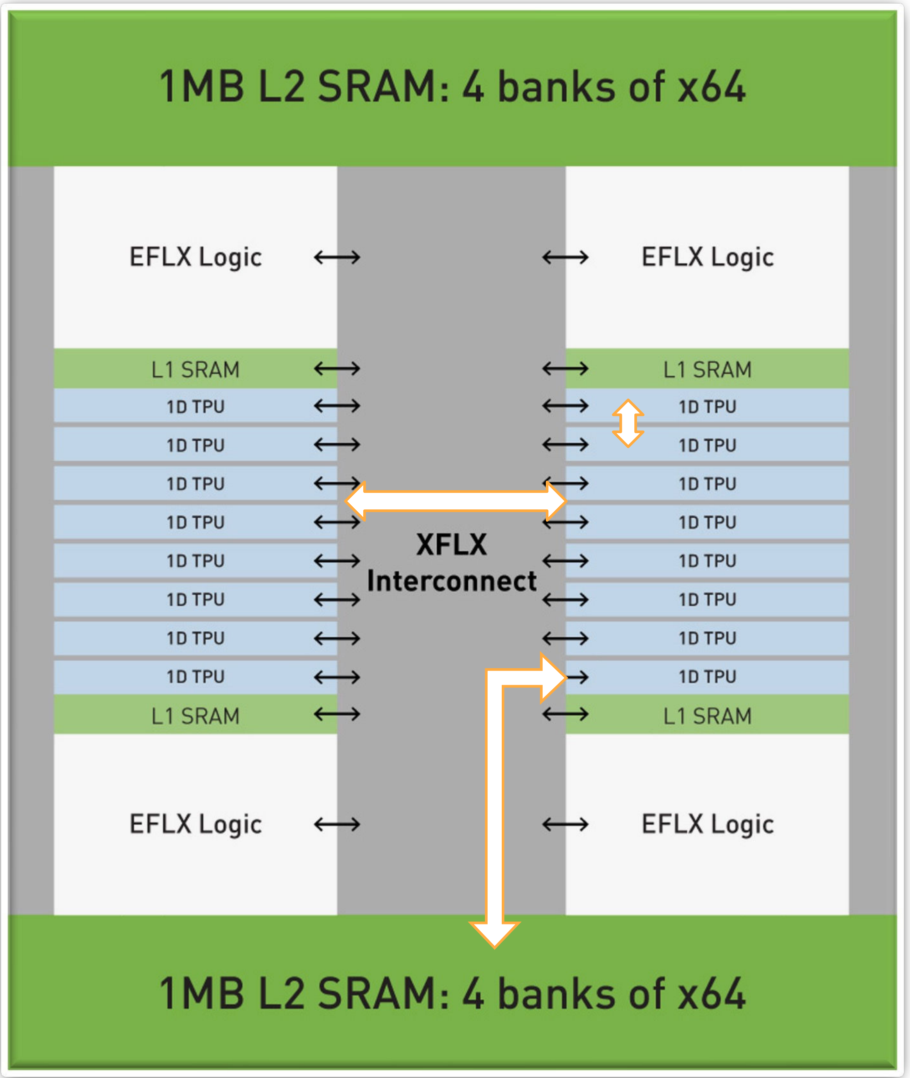
Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

DETR 2020 - Combines CNN Backbone with Transformer-based Detector



- Still uses ResNet or similar CNN backbone for feature extraction
- But, the transformer prediction head is fundamentally different:
  - Encoder will extract features across all "patches" to gather overall context
  - Decoder stages then makes prediction based on encoder results
  - Encoder/decoder computation is very different from CNNs, not suitable for traditional accelerators

Dynamic TPU: Flexible, balanced & memory-efficient



Efficient data access:
Each 1D TPU core can stream data from:

- Neighboring 1D TPU (dedicated)
- Any 1D TPU (via XFLX)
- L2 SRAM (via XFLX)
- DDR & System memory (via NoC)

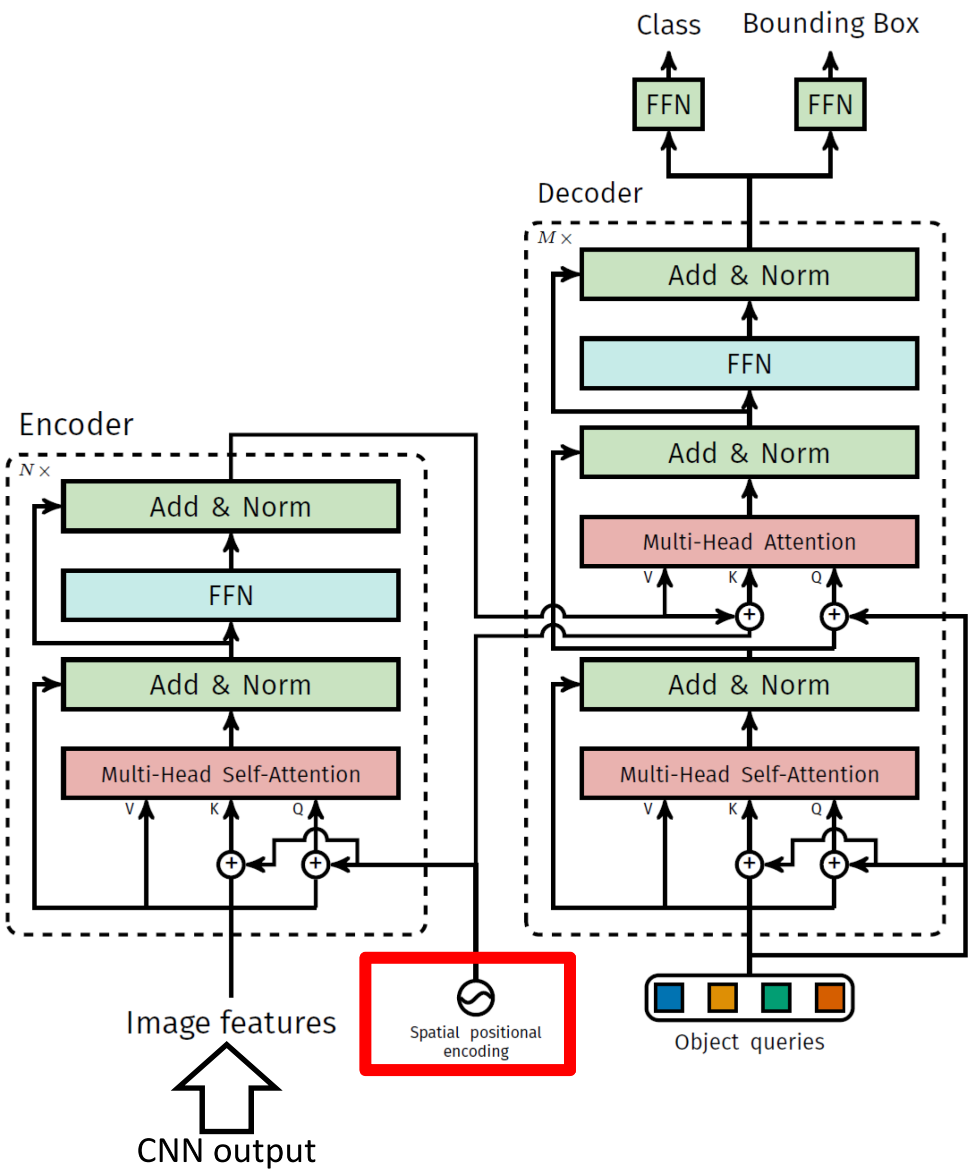
Flexible control and data path:

✓ “Future proof” compute, activations, and generic operators via EFLX
✓ X1 provides much more data bandwidth and manipulation vs NoC based AI engine
✓ X1’s flexibility include **mixed-precision support** – essential for supporting operators such as **transformers**

flexlogix

Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

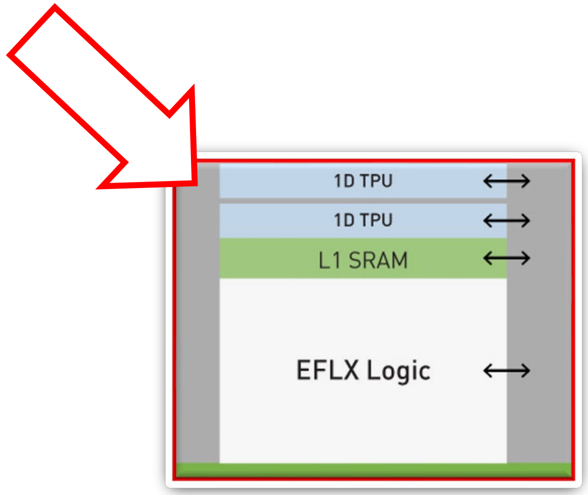
Diving into Vision Transformers (1)



First encoder stage is positional encoding:

- PE values are embedded into EFLX RAMs as ROM
- EFLX logic performs ROM look up "on the fly" prior to the V/K/Q attention head

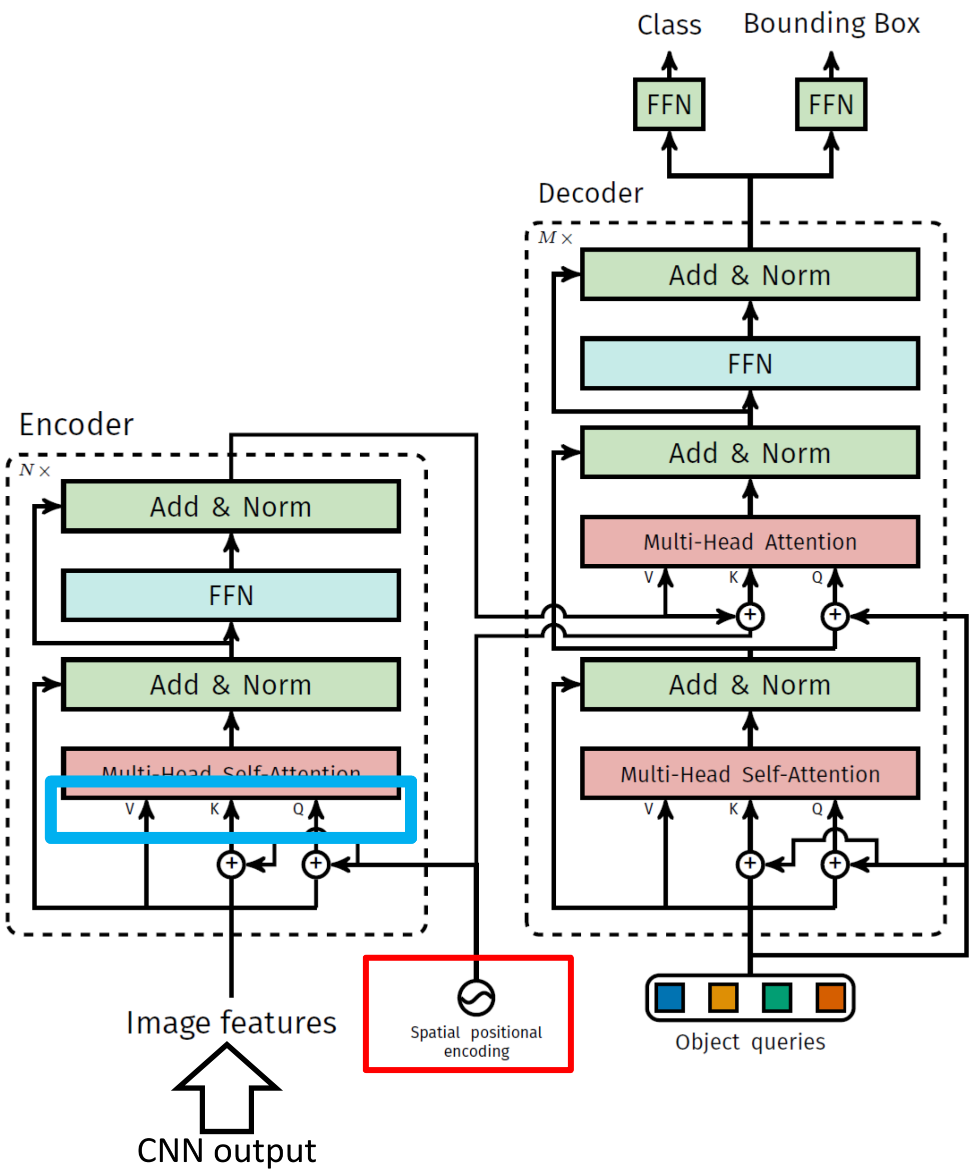
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$



flexlogix

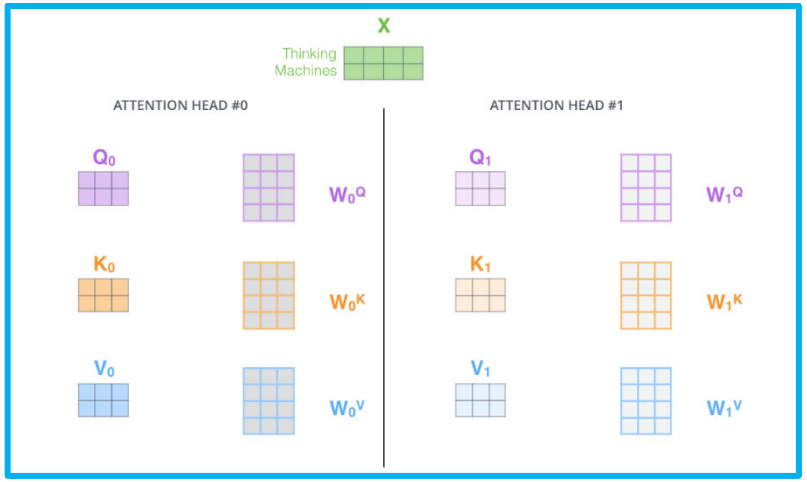
Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

Diving into Vision Transformers (2)



Second encoder stage is the multi-head attention layer

- Starts by multiplying input data with 3 separate matrices (Q, K, V) for each attention head
- Natively maps onto our 1D-TPU with (Q, K, V) as weights

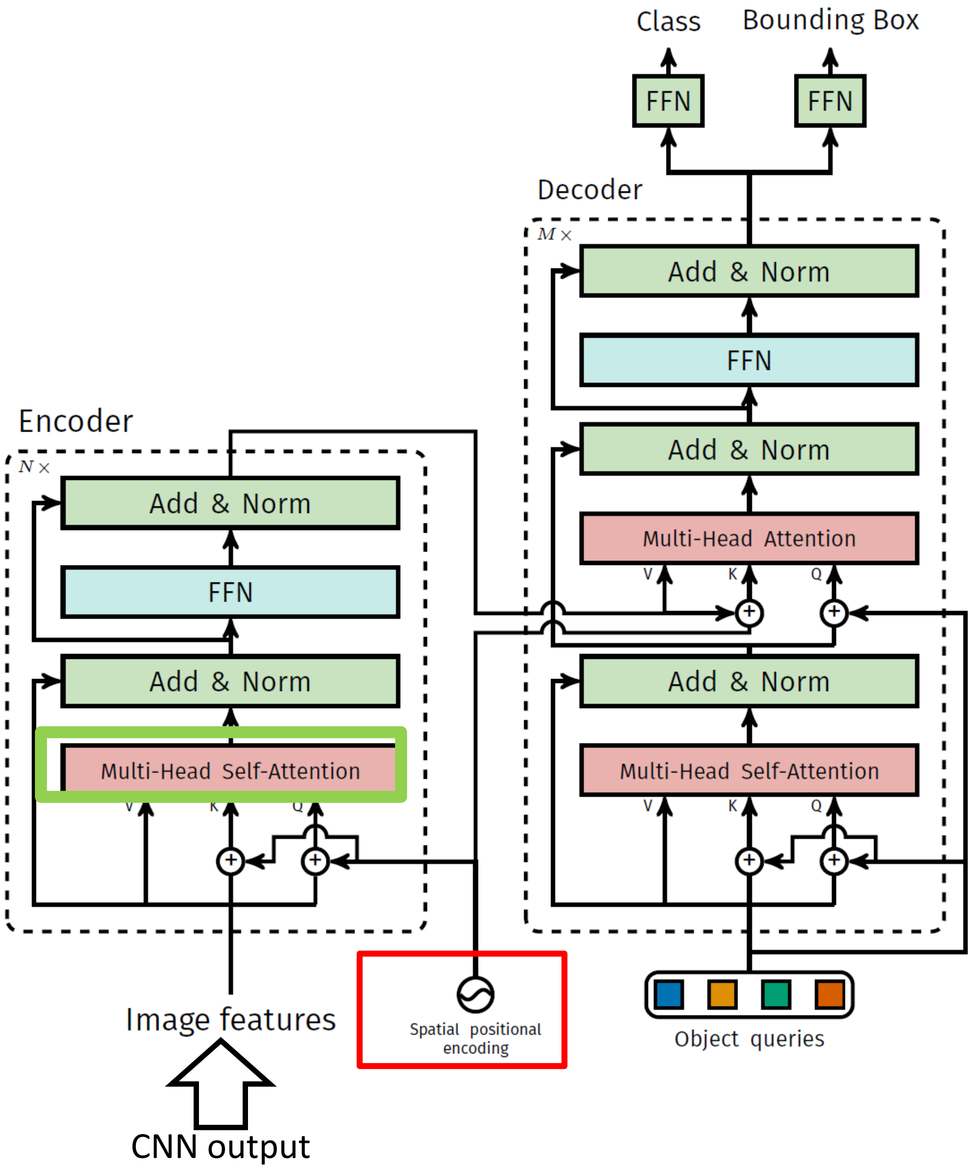


*https://jalammar.github.io/illustrated-transformer

flexlogix

Jeremy Roberson, Director of Inference SW, Flex Logix, September 2023

Diving into Vision Transformers (3)

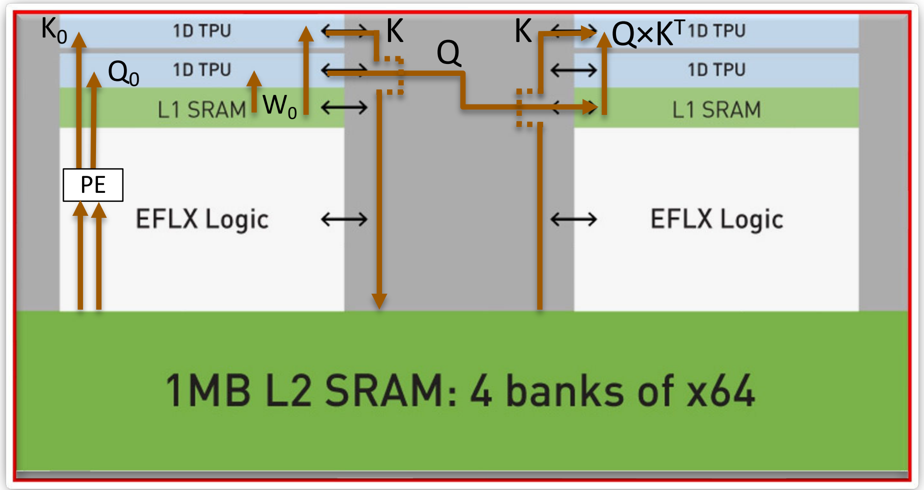


Main part of multi-head attention layer is a challenge on traditional edge accelerators:

- The (Q, K, V) for each matrix is now activation data

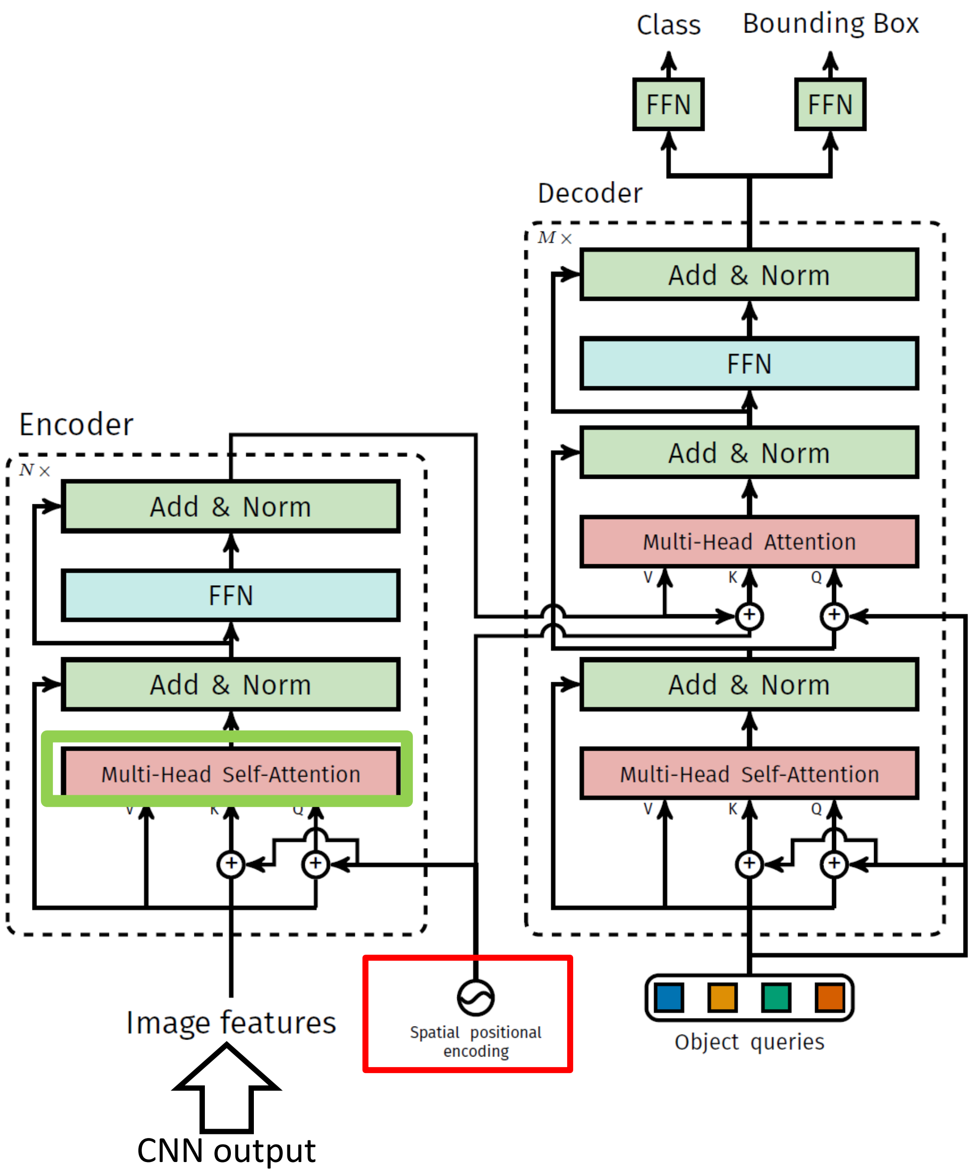
$Q \times K^T$ involves multiplying 2 activation data with **each other**

- Fortunately, X1 chip has dedicated path to load activation data into the weight memory



Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

Diving into Vision Transformers (4)

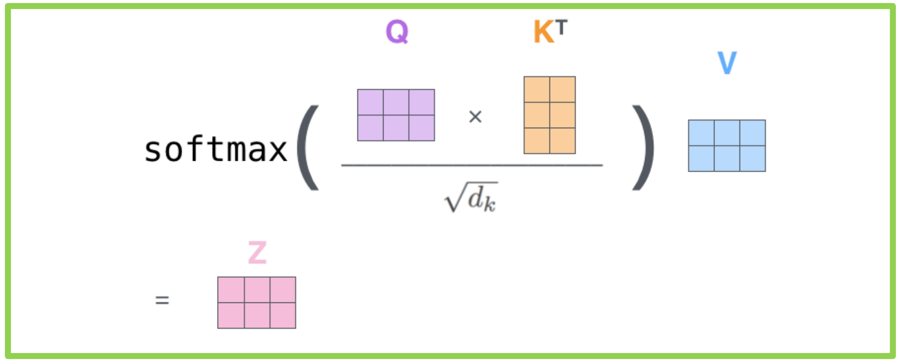


Softmax and norm are also challenging on int8 datapaths

- Fortunately, mixed-precision mode on X1 allows for z_i = Q×K^T result to convert to BF16 to execute exp(z_i)

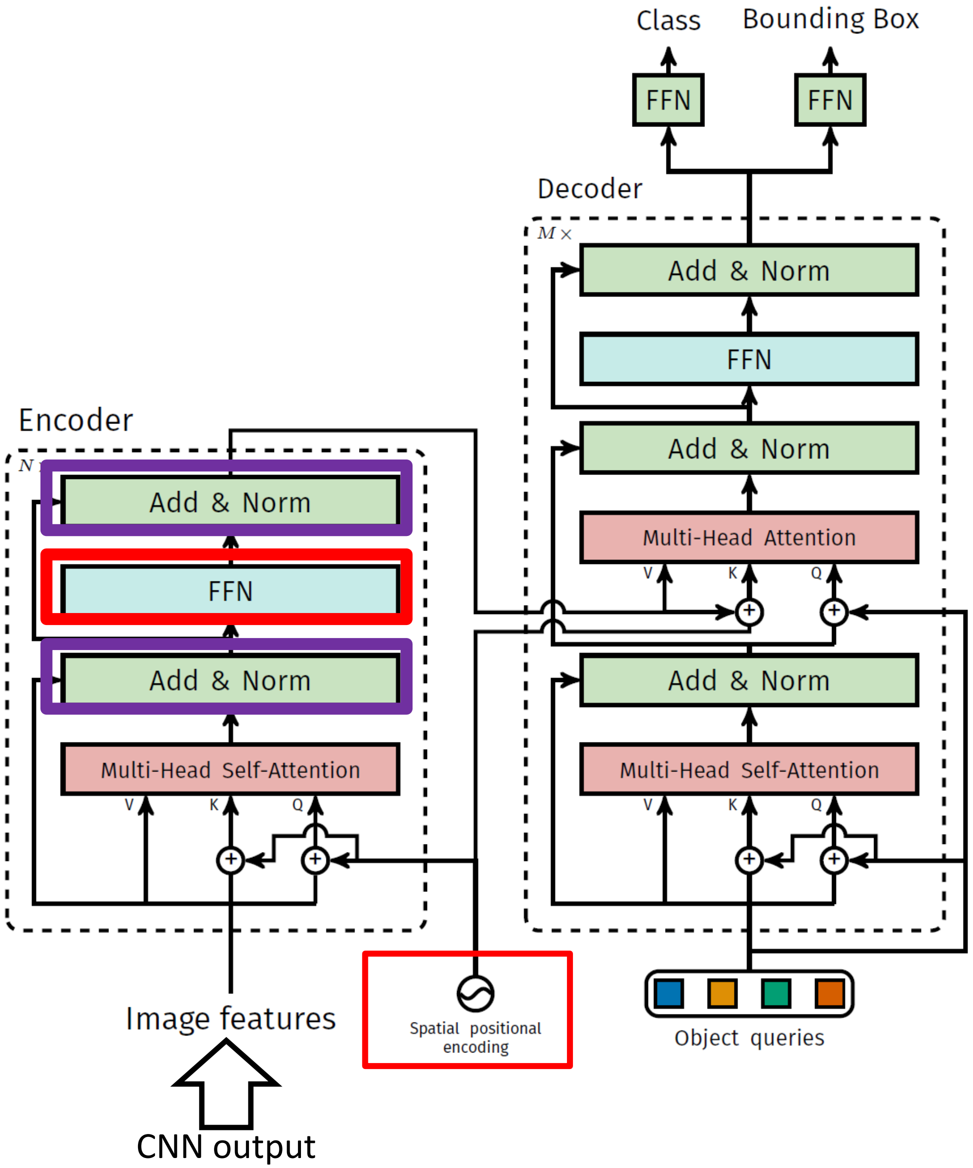
softmax $$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

- After normalization, softmax output is converted back to int
- Another **activation×activation** is performed natively on X1 to complete the multi-head attention layer



*https://jalammar.github.io/illustrated-transformer

Diving into Vision Transformers (5)



Add & Normalization follows the multi-head attention layer

- The original input is add with the attention output
- Normalization (esp. L2 norm) is best executed on BF16 due to its large dynamic range

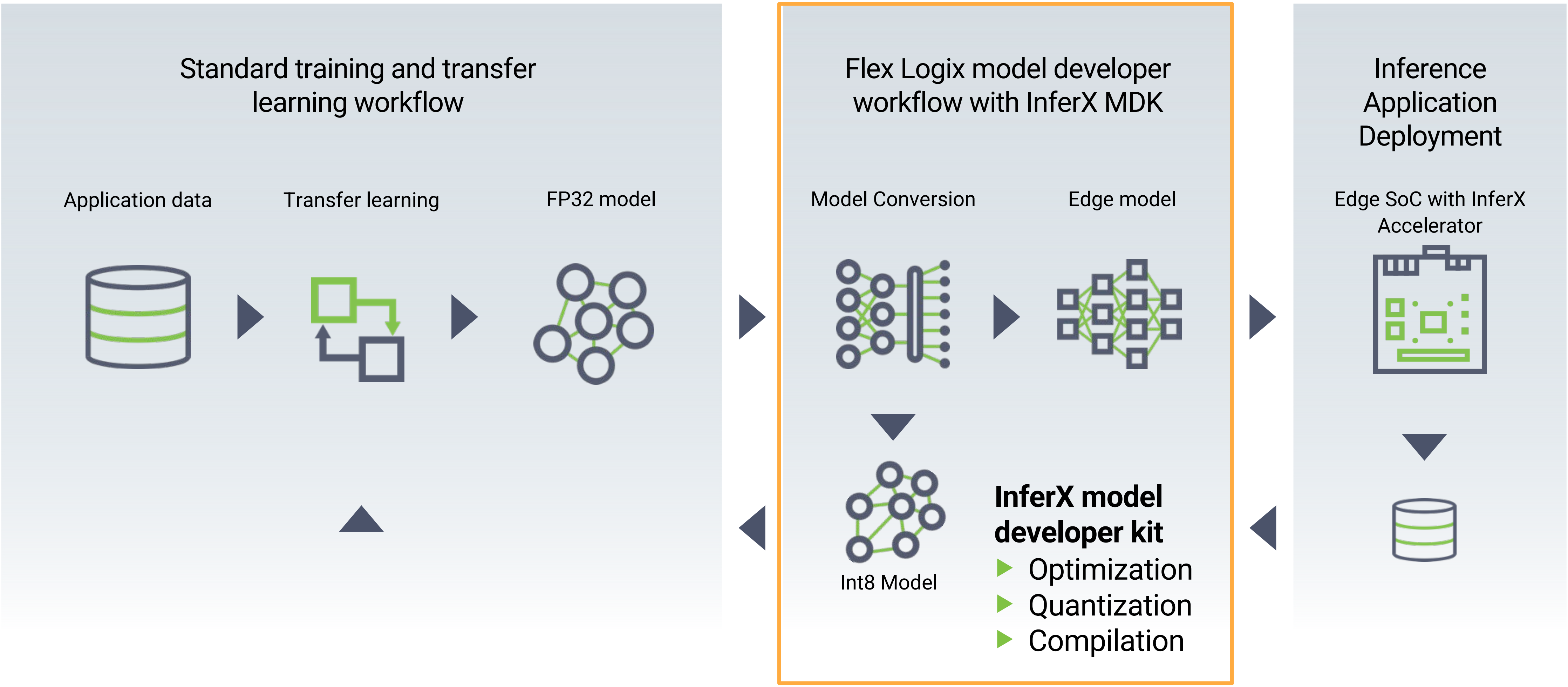
L1 norm $\|W\|_1 = \sum_{i}^{n} |\omega_i|$

Squared L2 norm $\|W\|_2^2 = \sum_{i}^{n} \omega_i^2$

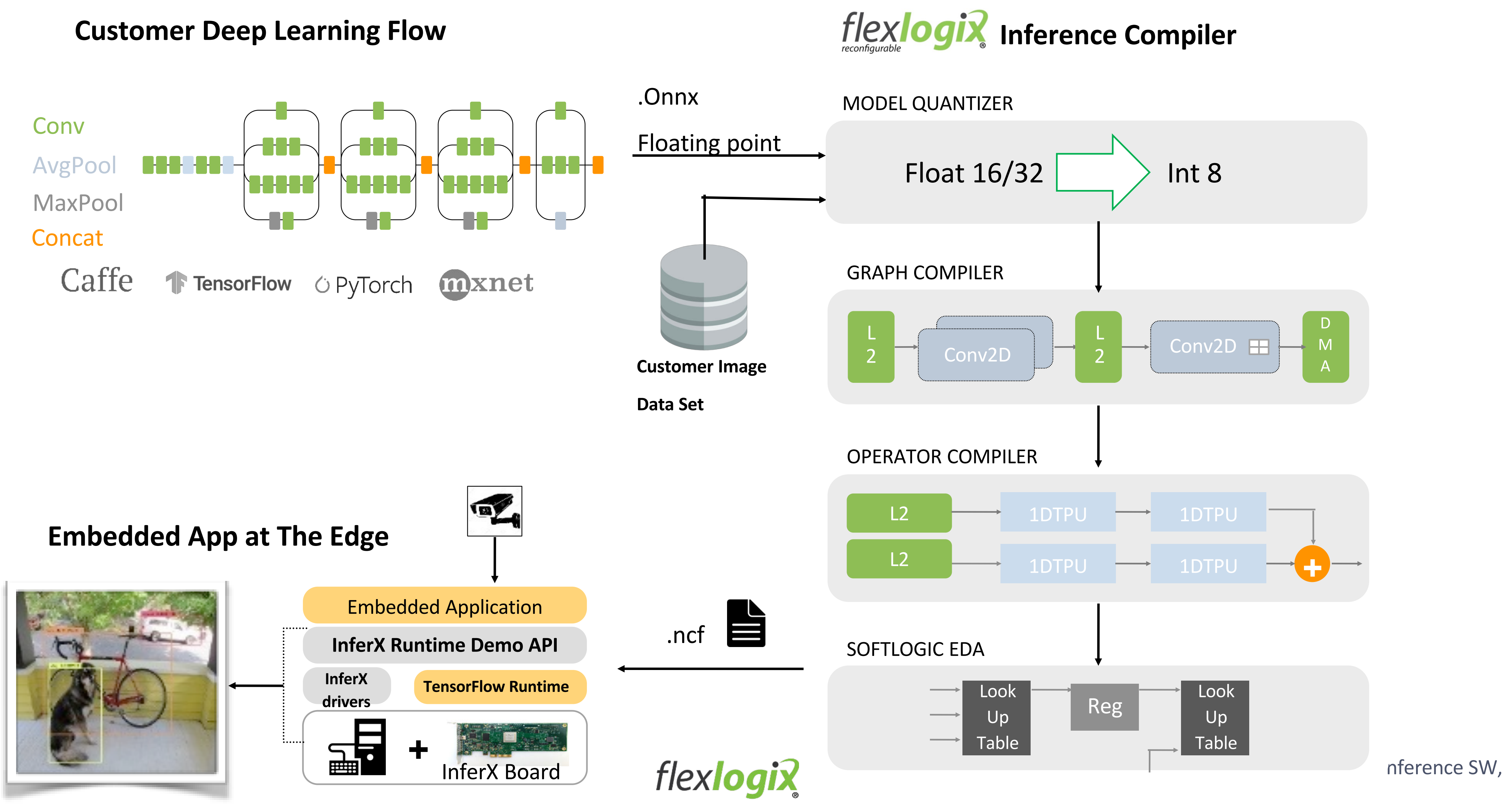
- Feed-forward Network (FFN) is a straight-forward GEMM operation before Add & Norm takes place again

*https://jalammar.github.io/illustrated-transformer

How to deploy InferX models in your system?

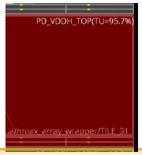


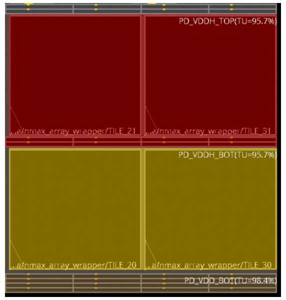
InferX compiler is available for evaluation Today

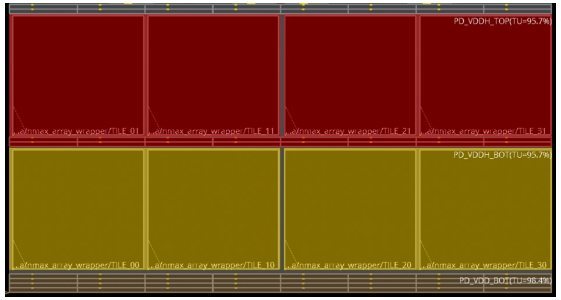


InferX IP family scales from a single tile to a large accelerator

InferX (N7)	Orin AGX 60W
16-128 DTOPs	138 DTOPS
1-4 LPDDR5 x32	LPDDR5 x256 (half for AI)







	16 TOPS 1 LPDDR5	32 TOPS 1 LPDDR5	64 TOPS 2 LPDDR5	128 TOPS 4 LPDDR5
DETR 2020 (Transformer) (1024×1024)	19 IPS	39 IPS	77 IPS	147 IPS
YOLOv5s (640×640) **Orin AGX 125 IPS**	200 IPS	330 IPS	716 IPS	1123 IPS
YOLOv5l6 (1280×1280)	12 IPS	19 IPS	**Orin AGX 31 IPS** 43 IPS	100 IPS
ResNet50 (1024x1024)	29 IPS	39 IPS	84 IPS	197 IPS

flexlogix®

Jeremy Roberson, Director of Inference SW,
Flex Logix, September 2023

Thank You

flex-logix.com

Flex Your Computing